DOCUMENT RESUME

ED 170 020 **PS 010 322**

AUTHOR Price, Gary G.
TITLE Standardized Achievement Tests for Young Children: An Analysis.
PUB DATE 30 Oct 78
NOTE 18p.; Paper presented at the Big Ten Symposium on Early Childhood Education (Madison, Wisconsin, October 30, 1978)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Criterion Referenced Tests; *Early Childhood Education; *Educational Assessment; *Educational Objectives; Educational Philosophy; Elementary School Curriculum; Minimum Competency Testing; Norm Referenced Tests; Preschool Curriculum; *Standardized Tests; *Testing Problems; Test Reliability; Test Validity

ABSTRACT

This article analyzes potential problems in using standardized achievement tests in early childhood education settings. The first problem discussed is whether the outcomes typically measured by standardized achievement tests are the outcomes about which early childhood educators are most concerned. Second, it is argued that a concern for test reliability must be balanced against whether the tests are appropriately specific or general and what kinds of skills they measure, etc. Third, the problem involved in measuring only intended program outcomes is noted. It is suggested that educators must be concerned that programs meet certain minimal standards (intended or not), that the programs optimize their intended goals, and that the tests should reflect these concerns. Fourth, more specific problems of testing are pointed out, i.e. (1) test items being more appropriate for some curricula than others, (2) items requiring more specialized knowledge than they were intended to test, (3) items being developmentally inappropriate, (4) preparation for the test taking time away from other valuable activities, and (5) the difficulty of testing for traits with delayed effects. Fifth, the difference between broad-domain and narrow-domain tests is explained, and the dangers of organizing the curriculum around these tests are noted. Suggestions are made on using tests and explaining the uses and limitations of tests to the public. (BH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

- ED170029

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Standardized Achievement Tests for Young Children: An Analysis

Paper presented at the
Big Ten Symposium on Early Childhood Education
Madison, Wisconsin
October 30, 1978

Gary G. Price
University of Wisconsin-Madison

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gary G. Price

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

PS 010322

"Now we are ready to begin the test. If you are not sure of an answer, mark the answer you think is right. If you change your answer, erase the mark you made and make a new mark. Are there any questions? Look at the picture of the zebra. Find the letter that you hear at the beginning of the word 'zebra.' Mark your answer."[1]

Standardized achievement tests, long familiar in the upper elementary grades and recently more common in the primary grades, are now entering the kindergarten. The routine administration of such tests is itself an educational practice. Like other educational practices, it has costs and presumed benefits. Before introducing this educational practice to the primary grades and kindergarten, one could fairly ask whether education in the upper elementary grades has improved since the introduction of the practice there. Such a question presupposes, perhaps wrongly, that a major purpose of achievement testing has been to improve education. In fact, the practice often has been motivated by other concerns -- such as meeting the requirements of funding agencies, responding to local political pressures, and preparing records of a kind that may one day be needed in court.

The present motivation for administering achievement tests to young children may be different, given the widespread public concern for "basic skills." Standardized achievement tests constitute convenient and reputedly legitimate operationalizations of "basic skills." They are therefore attractive as evaluation criteria to persons making decisions about educational programs for young children (and about personnel involved with those programs). The term "program" as used here should

be regarded as interchangeable with "preschool," "kindergarten," and "teacher." In other words, the present climate seems to be one in which achievement tests, regarded as indicators of the extent to which educational goals are being attained, may come to play a more active role in policy decisions and personnel decisions.

What unintended outcomes are likely to result from guiding educational activities with narrowly defined goals like those implicit in most achievement test batteries? What else, besides higher achievement test scores, should be the intended outcomes of education for young children? These are important questions that will lie behind my discussion, but I will focus more on tests and their interplay with the goals of early childhood education programs.

First, it is instructive to point out that most early childhood educators are in the habit of conceptualizing program goals in terms of outcomes. This is most evident in our criticisms of programs. In our criticisms, we fault programs for failing (in actuality or in our predictions) to produce the outcomes we value most. Alternatively, we fault programs for producing (in actuality or in our predictions) outcomes that we regard as undesirable.

Not only do early childhood educators conceptualize program goals in terms of outcomes, the outcomes are of a particular type. We almost invariably attend to outcomes in which the individual child is the locus. Achievement testing, which measures the individual child, is certainly consistent with that habit of designating the individual child as the locus of outcomes. Other loci are possible, and it would behoove early childhood educators to include them more explicitly in goal statements and in program criticisms. One alternative locus is the group. An example of a group outcome would be the prevalence of cooperative play

among the children. Another would be freedom from cliques. While early childhood educators certainly concern themselves with such sociological phenomena, our predominantly psychological perspective often leads us to highlight other goals.

Another locus of outcomes would be the environment itself, whose subtleties have been ignored by most educational evaluators and researchers (Schulman, 1970). We might consider the achievement of ethnic integration in a program to be a desirable outcome. Successfully maintaining a sanitary and nutritionally balanced meal program could be an outcome. Other, less easily operationalized program characteristics could be outcomes. Treating program characteristics as outcomes does not meet what most of us have in mind when we think of program outcomes. But the definition is not critical; the point is that evaluations of educational activities could be made without direct reference to outcomes.

The point may be clearer if I choose an extreme example. Imagine a preschool program where children are verbally abused as a routine matter. Imagine further that, as gauged by every available measure of the individual child -- social competence measures, achievement tests, health measures, and so on -- the children of that program seem better off than children in your favorite programs. Even without outcome measures as evidence, most early childhood professionals would want that program to be changed or shut down.

The evaluation of social institutions without measurement of outcomes is not new. As Herbert Zimiles (1977) has pointed out, museums, libraries, summer camps, and many other institutions are commonly evaluated without direct assessment of the beneficial outcomes they are believed to produce. Zimilies argues that, if we systematically examine the environment provided by an educational program, we can then use existing theory to make a

predictive assessment of the environment's potential for producing desired outcomes. The public might accept that view and so might practitioners, but psychologists and social scientists attempting such predictive assessments would often become mired in the discrepant predictions of competing theories.

Elliot Eisner (1975) and Dave Berliner (Note 1) have retreated one step further from assessment of outcomes by arguing that, in some situations, evaluators should function as connoisseurs who make judgements about programs without necessarily explicating any outcome-prediction basis for their judgements. An obvious criticism of this arrangement is that one could expect low inter-connoisseur agreement. Berliner would argue that the choice between disagreeing connoisseurs is often made in other contexts; the choice involves decisions that are more political than scientific. It is a choice that requires no standards other than those one implicitly uses when deciding who is a connoisseur -- that is, whose opinion one is willing to trust. Discrepancies between theoretical predictions, on the other hand, will continue to exist for various reasons. One reason is that the very phenomena that theories interpret can undergo change (Cronbach, 1975). Another reason is that theoretical development is hindered by measurement methodology, which is also why many desired outcomes are not measureable. Current measurement methodology is primarily suited for low-inference constructs,[2] whereas high-inference constructs are often of more interest -- theoretically and educationally.

In view of how limited we are in our methods of measuring desirable outcomes, one might expect widespread enthusiasm for standardized achievement tests, with their ease of use and superior reliability. But those attributes alone are not sufficient to commend the use of a measure in educational evaluations. Bathroom scales and measuring tapes are easier

to use with young children and even more reliable, but few of us would think to use them in evaluating a program for young children. This absurd example should make it clear that ease of use and reliability -- properties of standardized achievement tests that are often extolled -- mean little in isolation. But it should also make it clear that standardized achievement tests have some other properties not possessed by bathroom scales and measuring tapes. Standardized achievement tests are popular evaluation tools because they do have some degree of validity as measures of trait-like psycho-educational constructs that are integral to the educational priorities of many persons.

There is more to say about the match between achievement tests and program priorities, but first I would like to clarify the role of reliability in measures of individuals. Standardized achievement tests certainly qualify as measures of individuals, but so do all instruments meant to classify individuals along an ordered scale that runs from low to high with respect to the characteristic being measured. Instruments meant to measure motor coordination, self-confidence, communication skill, dramatic ability, emotional stability, and so on, would fit this definition. The point I want to make is that, for some purposes, reliability considerably lower than that of standardized achievement tests is fine. It all depends on what kind of decision is being made and the role of individuals' scores in that decision. If scores of individuals are used as a basis of long-term placement or selection, then high reliability is very important (as is also the reasonableness and justice of basing the placement or selection on the characteristic measured). On the other hand, if scores of children are averaged and the decision to be made requires only a rough estimate of the group mean, then no harm is done by using unreliable measures. The reason for this is that children who receive uncharacteristically

high scores on the measure will tend to be offset by children who receive uncharacteristically low scores, leaving the group mean reasonably close to what it would be if the measurement error could somehow be removed. This remark about the stability of the group mean gets truer as the group gets larger, in accord with the law of large numbers. Gross misassessments of group means are therefore unlikely if a measure's only fault is unreliability.

The catch in what I have said is that unreliability does little harm if the decision to be made requires only a rough estimate of the group mean. Sometimes decisions do require a precise estimate of the group mean. Sometimes decisions may require information other than the group mean. A crucial question that is often ignored is whether the group mean (regardless of the precision with which it is estimated) is an appropriate kind of information on which to base a decision. I suspect that in many situations where a rough estimate of the group mean would have sufficed, moderately reliable outcome measures that reflected the intended goals of a program or the suspected ill effects of that program unfortunately were left out of evaluators' batteries and policy makers' thinking.

The position I have advocated here is one of balancing the appropriateness of an outcome measure against its unreliability. If a measure is highly appropriate but not very reliable, it should still be used -- so long as a rough estimate of its mean would be useful in making a program decision. Evaluators, typically trained to do research directed at generalizations about parameters, too often are prone to conclude that information is not useful if it is not precise. The position advocated here also dictates that an inappropriate measure should not be used, no matter how reliable it is. A presupposition of this position is that judgements can be made about the appropriateness of candidate outcome

measures.

One way to gauge the appropriateness of an outcome measure is to ask whether it measures the intended outcomes of the program being evaluated. Persons involved in a program for young children would naturally bristle at having their efforts evaluated on the basis of criteria that are alien to their primary concerns. Tom Carpenter and I have argued elsewhere that, when evaluations are based on outcome measures which have not been chosen to reflect the goals of the program being evaluated, the priorities of the program are thereby rejected a priori and replaced with the priorities implicit in the set of outcome measures (Price & Carpenter, 1978).

Since making that argument, I have become aware of its inadequacies. If one takes the position that congruence with program goals is the only gauge of an outcome measure's appropriateness, then one makes the program accountable for its stated priorities, and unaccountable for anything else. Remember the earlier example of a program in which children are verbally abused. Measures of the suspected ill effects of a program seem every bit as appropriate as measures of its intended outcomes.

The assertion that programs should be accountable for outcomes other than those included in their stated priorities presupposes that there are some outcomes whose attainment or avoidance ought to be common to all programs. Determining what ought to be common priorities of this sort is ethically and politically problematic, but nonetheless inherent to educational decisions. Early childhood professionals concerned about the evaluation of programs with achievement tests would not deny that some common priorities should exist. Their main concern is that the agents who govern policy and personnel changes -- be those agents administrators or the public -- are setting skewed priorities by directing attention to standardized achievement tests.

Is there no special status accorded to the intended outcomes of a program? I would argue that the intended outcomes of a program do deserve special status. Consider the difference between optimization and meeting minimal standards. It is reasonable to expect an optimal showing from a program on its intended outcomes, assuming that they can be measured. On other measures, the program should meet, but not necessarily exceed, minimal standards. This distinction loses its usefulness if persons enamored with some particular outcome -- such as early reading -- are permitted to convince the public that their pet outcome is one which all programs should optimize. Early childhood educators can safeguard such skewing of priorities by informing the public of the multiple concerns in early childhood education and by responsibly altering the expectations that people have about specific types of achievement. That seems to me a better strategy for concerned early childhood educators than categorically opposing the use of achievement tests.

In relating standardized achievement tests to program goals, it is useful to keep in mind that tests are really collections of separate items, and that evaluations based on tests really start at the level of the test item. In an analysis of the Follow Through evaluation, for example, Gene Glass (Note 2) demonstrated that the apparent superiority of the Direct Instruction model on a math achievement test stemmed from a few items in which arithmetic problems followed a format familiar to the Direct Instruction children and unfamiliar to other children. Items like those, in the vocabulary of program evaluators, are curriculum reactive.

The concept of curriculum reactivity grows out of the more general concept of reactive measurement effects (Webb, Campbell, Schwartz, & Sechrest, 1966). Reactivity refers to the tendency of a measurement procedure to misrepresent the observed phenomenon, typically by altering

the phenomenon itself through the intrusiveness of the measurement procedure. A test is also reactive if the test score is substantially weighted on an ability or knowledge that is tangential to the construct which the test is reputed to measure. Thus, a test is curriculum reactive if it discriminatingly rewards or penalizes students by testing ancillary knowledge peculiar to certain curriculums. The most obvious form of cultural bias in a test is a special case of curriculum reactivity. Children who have experienced a particular cultural environment (curriculum writ large) may not have acquired information or skills that are essential for success on some items, yet irrelevant to the construct being measured.

Another way to gauge the appropriateness of an outcome measure is by its developmental appropriateness. If an achievement test contains items that young children are developmentally unequipped to tackle, then it is not an appropriate outcome measure to use with those children, regardless of the program they have experienced. Some of the older achievement tests which were designed for upper elementary school children and minimally adapted for use with younger children could be attacked as developmentally inappropriate. Some of their obvious faults are answer forms that require more dexterity than one can expect of many young children and directions that are unduly long and complex. In some recently developed standardized achievement tests, it should be noted that the publishers have focused on narrow age ranges, and have made an obvious effort to tailor the format of the test to the special requirements of young children.

Norm-referenced achievement tests, when standardized on a narrow age range, seem serendipitously assured of being developmentally appropriate. The process of age-norming a test ensures that most of the items included in it are ones which children of that age have a reasonable likelihood of performing successfully. If a test were to contain very many items

beyond the reach of the young children for whom it was intended, it simply would lack the psychometric qualities traditionally sought by test publishers. In that respect, the age-norming of standardized achievement tests is an avenue by which young children can, through their test performance, guide test publishers away from unrealistic expectations. This inherent safeguard, which fosters the developmental appropriateness of norm-referenced achievement tests, is not inherent in criterion-referenced or competency-based tests.

Criterion-referenced tests can certainly be developmentally appropriate, but their fitness for children of a particular age depends more fully on the wisdom of the adults who choose the criteria. Performance on a criterion-referenced test is compared to an absolute standard, thereby removing the competitive element of norm-referencing. But the acclaimed differences between norm-referenced and criterion-referenced tests often vanish when comparisons are made at the level of individual test items; many items would be suited for either type of test. Ultimately, both types of test must be judged on the basis of whether the specific tasks that they present to children are developmentally appropriate, educationally reasonable, and not artificially difficult.

There is another kind of developmental appropriateness, which is more subtle. A test can require information and skills achievable by young children, but achievable only through all-consuming efforts that supplant activities more critical for children of preschool and kindergarten age. This kind of developmental inappropriateness is actually an inappropriateness of program priorities, which can be embodied in achievement tests. The assertion that some program goal is incompatible with other, more critical experiences rests on the premise that early childhood is a critical period during which it is especially important that children have certain kinds

of experiences, a premise held by most early childhood educators.

Earlier I discussed the principle of expecting a program to meet minimal standards on some outcome measures, and to give an optimal showing on its intended outcomes. But what about the many programs in early childhood education whose intended outcomes are impractical to measure? Though such programs may meet any number of minimal standards, advocates of such programs will be under particular pressure to explicate why they expect that the environment they provide for children will make certain valued but unmeasureable outcomes likely. If one cannot show the results that make a program worthwhile, one should be prepared to present the reasons for believing that it is worthwhile.

Early childhood education as a field is peculiar in that it is often concerned with long range goals and delayed effects, neither of which are gauged adequately in a short time frame. It is especially common to hear early childhood educators recommend against a practice because it is believed to have an undesirable delayed effect. While few early childhood educators today use the explanatory vocabulary of psychoanalysis, many take for granted the general type of phenomenon that Freud built his theory to explain -- the phenomenon of delayed response to early childhood experience. Recent press coverage of diseases with long latent phases (like cancer caused by exposure to asbestos) has probably made the public more appreciative of "delayed effects" arguments. So early childhood educators would seem to be in a position to make the public more cognizant of remote goals which the public, too, would value, but which are not assessed easily.

Even many traditional academic goals are gauged poorly in a short time frame. Generally, abilities and knowledge that young children are asked to demonstrate by performing tasks on achievement tests are isolated

prerequisites of complex, socially valued competencies that are seldom (if ever) fully developed in young children.[3] Early acquisition of the prerequisites provides no guarantee that the complex competencies themselves will be acquired sooner or with more refinement. Consequently, short-run measurement of prerequisite knowledge can underestimate the prospects of some children, especially children in programs where those prerequisites are not salient. While it may be years before measurable knowledge of this kind becomes critical as a prerequisite, it often functions well before than as a selection criterion in the social assorting done byteachers and peers (Rist, 1970).

To return more particularly to standardized achievement tests, I would like to point out a trend toward segmenting tests and reporting performance in terms of narrower domains. Heretofore, the common practice in achievement testing has been to report performance in terms of labels that correspond to traditional curriculum areas, particularly with respect to "reading achievement" and "math achievement" -- with token attention to (less reliable) subtest scores like "getting the main idea" and "computational skill." At least one of the new tests has dropped the use of the overall "reading achievement" score, and instead reports scores in domains like auditory discrimination, receptive vocabulary, comprehension of oral language, letter and numeral discrimination, letter and numeral recognition, the ability to orally describe common objects, and so on. In light of the recent history of using IQ tests as achievement tests in evaluations of programs for young children, this certainly constitutes a shift from broad domains to narrow ones.

The use of broad-domain tests to evaluate programs for young children has been widely criticized (McClelland, 1973; Zigler & Trickett, 1978). Despite the faults of broad-domain tests, their use in evaluations often spared early childhood programs from oversimplified criticism of curriculum,

since no simple change in a program could be expected to change scores very much on a broad-domain test. The constructs measured by broad-domain tests like IQ tests and Scholastic Ability tests pervade performances in different curriculum areas, so such constructs would not be affected greatly by instructional efforts in any particular curriculum area. Changes provoked in a program by low IQ scores or low Scholastic Ability scores usually took the form of increased opportunities for problem-solving, classifying, etc., which left the prior balance between curriculum areas relatively unperturbed. Expressive arts, opportunities to develop social skills, and other activities mistakenly regarded as "less cognitive" sometimes suffered, but no more so with broad-domain tests than with other achievement tests.

The shift to achievement tests whose domains are labeled to correspond with traditional curriculum areas makes it much more likely that test scores will be used as a basis for program changes. The reduction of achievement in curriculum areas like reading to a single scale leads the unsophisticated to entertain simplistic program changes that suppose instruction in a curriculum area to be a rather homogeneous activity whose efficacy depends mainly on the time devoted to it. A common response for a teacher whose children made a poor showing on a test labeled "reading achievement" would be to spend more time on reading, using familiar approaches. If, previously, the teacher had been systematically avoiding reading instruction, then increased time and effort would make better readers of some of the children -- no matter what instructional approach was used. Assuming that the teacher had already been devoting ample time to reading or prereading activities, more of the same instructional approach would probably result in increased frustration for the children who did not resonate to that approach and diminishing returns for the children who did.

In that respect, achievement tests bearing the same label as traditional curriculum areas provide little useful information about how to change a program.

Achievement tests with even narrower domains like auditory discrimination, comprehension or oral language, etc., spare early childhood programs from oversimplified criticism, simply because they supplant a single outcome measure with multiple outcome measures. Such tests also provide more useful information about how to change a program, since they suggest which prerequisite skills need more attention. Tests of this kind are likely to be accepted and used by teachers if their instructional usefulness is readily apparent. No matter how instructionally useful such tests might be, however, one may still question whether the curriculum goals they exemplify ought to be high priority ones. In that respect, instructionally useful tests present a seductive pressure for curriculum change, since program modifications motivated by them can shift the de facto curriculum priorities of a program even when no such shift is intended.

In conclusion, early childhood educators should actively forestall simplistic thinking about curriculum by informing and reminding all parties concerned of the variety of goals that deserve attention in programs for young children. The main issue -- balance in program priorities -- is only obscured by condemning norm-referenced standardized achievement tests or by advocating criterion-referenced tests. Some antiquated tests and badly formed items excepted, standardized achievement tests do a comparatively good job of measuring what they claim to measure; their primary weakness is in what they do not measure. The task domains included in standardized achievement tests are a common denominator of curriculum goals -- a common donominator of practical-to-measure curriculum goals. Neither being part of that common denominator nor being practical-to-measure should be equated with importance.

Footnotes

1. These instructions were taken from the California Achievement Test. Item content was changed so as not to compromise the content of the test.

2. Trait models that explain differences between the net performances or predilections of individuals admittedly involve higher-inference constructs those used by behaviorists. But traits differ from behaviorists' constructs primarily insofar as the trait is regarded as an aggregation (e.g., arithmetic mean) of all the low-inference observations germane to the trait that hypothetically could be made of the individual. The same may be said of latent trait models in the factor analytic tradition; those models rest on the postulation that test scores or even item scores are sums of multiple low-inference observations whose respective aggregations define the latent traits.

3. Abilities and knowledge are identified as prerequisites when, according to a formal analysis of a task, they appear to be logically necessary. A common fallacy of the task analyst is to presume that the formal categories used in the task analysis correspond to the actual cognitive processes, needs, and perceptions of persons learning how to do the task.

References

Cronbach, L. J. Beyond the two disciplines of scientific psychology. American Psychologist, 1975, 30, 116-127.

Eisner, E. W. The future of the secondary school: A viewpoint. Curriculum Theory Network, 1975, 5, 127-138.

McClelland, D. C. Testing for competence rather than for "intelligence." American Psychologist, 1973, 28, 1-14.

Price, G. G., & Carpenter, T. P. Review of On further examination: Report of the advisory panel on the Scholastic Aptitude Test score decline. Journal for Research in Mathematics Education, 1978, 9, 155-160.

Rist, R. C. Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. Harvard Educational Review, 1970, 40, 411-451.

Shulman, L. Reconstruction of educational research. Review of Educational Research, 1970, 40, 374-375.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive measures: Nonreactive research in the social sciences. Chicago: Rand McNally, 1966.

Zigler, E., & Trickett, P. K. IQ, social competence, and evaluation of early childhood intervention programs. American Psychologist, 1978, 33, 789-798.

Zimiles, H. A radical and regressive solution to the problem of evaluation. In L. G. Katz (Ed.), Current topics in early childhood education. Norwood, New Jersey: Ablex Publishing Corporation, 1977.

Reference Notes

Berliner, D. C. "Clinical" techniques required for close and constant classroom observation. Paper presented at meeting of the American Educational Research Association, Toronto, March 1978.

Glass, G. V. Remarks made publically at the meeting of the American Educational Research Association, Toronto, March 1978. Also see Glass, G. V. Minimum competence and incompetence in Florida. Phi Delta Kappan, 1978, 59, 602-605.